На правах рукописи

Циликов Илья Сергеевич

РАЗРАБОТКА МОДЕЛИ ПРЕДСТАВЛЕНИЯ, МЕТОДОВ И АЛГОРИТМОВ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ТЕКСТА С ЦЕЛЬЮ ЕГО ФОРМАЛИЗАЦИИ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Специальность 05.13.01 – Системный анализ, управление и обработка информации (в технических системах)

А В Т О Р Е Ф Е Р А Т
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2010

Работа выполнена на кафедре автоматизированных систем обработки информации и управления Санкт-Петербургского государственного электротехнического университета им. В.И.Ульянова (Ленина) (ЛЭТИ)

Научный руководитель: доктор технических наук, профессор
Борис Яковлевич Советов

Официальные оппоненты: доктор технических наук, профессор
Лукомский Юрий Александрович,

кандидат технических наук, доцент
Раков Игорь Васильевич

Ведущая организация: Санкт-Петербургский государственный университет аэрокосмического приборостроения

Защита диссертации состоится «___» _____2010 г. в ___ часов на заседании совета по защите докторских и кандидатских диссертаций Д212.238.07 Санкт-Петербургского государственного электротехнического университета «ЛЭТИ» им. В.И.Ульянова (Ленина) по адресу: 197376, Санкт-Петербург, ул. Профессора Попова, д. 5.

С диссертацией можно ознакомиться в библиотеке университета.

Автореферат разослан «___» _____ 2010 г.

Ученый секретарь
совета по защите докторских и
кандидатских диссертаций Д212.238.07
кандидат технических наук,
доцент В. В. Цехановский

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. В настоящее время история развития информационных систем, т. е. систем, предназначенных для хранения и обработки информации с использованием ЭВМ, насчитывает уже более полувека. Еще относительно недавно в ходу были перфораторы в качестве устройств ввода данных, листинги в виде рулонов бумаги длиной порой до нескольких метров в качестве носителя результатов машинной обработки, недельные, либо месячные временные интервалы — в качестве нормативных сроков обработки информации. В последнее десятилетие прошлого века ситуация претерпела качественные изменения. Основу информационной системы в настоящее время составляют: база данных, как правило, реляционного типа, поддерживающая доступ на основе стандарта SQL, программные средства, обеспечивающие логику обработки данных, и интерфейс пользователя.

Применение баз данных благодаря специальным методам хранения и представления данных и соответствующим алгоритмам оперирования ими позволяет обеспечивать высокую производительность информационных систем, а наличие единого стандарта доступа к данным обеспечивает высокую эффективность их разработки и функционирования. Но с другой стороны применение баз данных требует специальной процедуры ввода данных, и если исходная информация представлена в виде неструктурированного естественно-языкового текста, то эта процедура становится весьма трудоёмкой, в виду чего становится актуальной задача автоматизации этой процедуры. Эта задача требует применения методов интеллектуальной обработки текста, которые активно развиваются в настоящее время. Хотя существующие на настоящее время методы интеллектуальной обработки текста не способны оценивать его структурированность в той степени, в какой эта характеристика текста отражается в человеческом восприятии, что не позволяет создавать полностью автоматические системы ввода данных, а кроме того производительность вычислительной техники на настоящее время остаётся всё ещё недостаточной для эффективной работы многих методов интеллектуальной обработки текста, тем не менее применение частично автоматизированных систем может существенно сократить трудоёмкость процедуры ввода данных, что обусловливает актуальность задачи разработки этих автоматизированных систем.

В области интеллектуальной обработки текста первым значительным успехом было появление контекстно-свободных грамматик Н. Хомского. В нашей стране большее распространение получила модель "смысл-текст" И. А. Мельчука. Возможные доработки и модификации этой модели были предложены Ю. Д. Апресяном, а также Е. В. Падучевой. В практическую реализацию систем интеллектуальной обработки текста, основанных на этой модели, большой вклад внесли А. В. Сокирко, П. В. Толпегин, И. М. Ножов, их предшественниками в этой работе были Н. Н. Леонтьева, С. Л. Никогосов, И. М. Кудряшова, О. Б. Малевич.

Развитие Internet'a потребовало широкого применения других методов интеллектуальной обработки текста, в первую очередь методов информационного поиска. Первый метод информационного поиска был предложен К. Муром в 1948 году, сначала его применение ограничивалось обеспечением доступа к книгам, журналам и другим документам в университетах и библиотеках. Первая поисковая система для Internet'a разработана М. Грэем из Массачусетского технологического института в 1993 году. Ранее в 1988 году С. Диэрвестером был предложен латентно-семантический анализ, основанный на теории сингулярного разложения, разработанной Дж. Сильвестром в 1889 году. Также в качестве одного из методов интеллектуальной обработки текста стал активно использоваться кластерный анализ, впервые предложенный Р. Трионом в 1939 году.

Тем не менее все эти подходы к интеллектуальной обработке текста не могли обеспечить качество решения различных задач, адекватное восприятию естественно-языковых текстов человеком. Одной из попыток достичь более высокого качества интеллектуальной обработки текста является начатый в США в 90-е годы прошлого века проект «Микрокосмос», работа над которым продолжается в настоящее время. Этот проект ориентирован преимущественно на решение задачи машинного перевода и основные его наработки касаются английского и испанского языков. Среди работ, выполненных в нашей стране, можно отметить семантический анализатор, разработанный В. А. Тузовым, а также разрабатываемый в настоящее время в Санкт-Петербургском институте лингвистических исследований открытый лингвистический процессор. Ещё один подход к интеллектуальной обработке текста предложен В. А. Фомичёвым.

Сложность применения перечисленных более новых подходов к интеллектуальной обработке текста для построения автоматизированной системы ввода данных в информационные системы с формализованной

структурой документа состоит в том, что не существует ни одной завершённой, общедоступной и практически применимой реализации какого-либо из этих подходов для русского языка. В связи с этим предлагается основывать интеллектуальную обработку текста на модели "смысл-текст" И. А. Мельчука, сочетая с элементами подходов, появившихся в связи с развитием Internet'a.

Объектом исследования являются информационные системы, использующие в своих данных естественно-языковый текст и использующие его формализованную структуру.

Предметом исследования являются модели представления естественно-языкового текста и алгоритмы для его формализованного структурирования.

Цель и задачи исследования. Основной целью представленной диссертации является разработка модели представления, методов и алгоритмов интеллектуальной обработки текста с целью его формализации в информационных системах. При этом решаются следующие задачи:

1. Разработать модель представления естественно-языкового текста на основе семантической сети для его интеллектуальной обработки текста с целью формализации в информационных системах
2. Разработать алгоритм структурирования естественно-языкового текста для его формализации в информационных системах в соответствии с такими критериями структурированности текста, как наличие иерархического оглавления, наличие заголовков у каждого из разделов оглавления, семантическая связанность внутри каждого из разделов.
3. Разработать методы и алгоритм интеллектуальной обработки текста на основе иерархической и бинарной кластеризации семантической сети для формирования иерархического оглавления естественно-языкового текста.
4. Разработать алгоритм интеллектуальной обработки текста на основе методов квазиреферирования для формирования заголовков у каждого из разделов оглавления.
5. Разработать алгоритм интеллектуальной обработки текста на основе исчисления предикатов для обеспечения семантической связанности внутри каждого из разделов оглавления.

Методы исследования. Для проведения исследований были использованы методы графематического, морфологического, синтаксического и первичного семантического анализа естественно-языковых текстов, методы иерархической и бинарной кластеризации, матричные вычисления, методы квазиреферирования, исчисление предикатов.

Основные положения, выносимые на защиту:

- Модель представления естественно-языкового текста на основе семантической сети.
- Алгоритм структурирования естественно-языкового текста для его формализации в информационных системах.
- Методы и алгоритм для формирования иерархического оглавления естественно-языкового текста.
- Алгоритм для формирования заголовков у каждого из разделов оглавления.
- Алгоритм для обеспечения семантической связанности внутри каждого из разделов оглавления.

Научная новизна работы.

- Предложена модель представления естественно-языкового текста, базирующаяся на модели «смысл-текст» в виде семантической сети, отличающаяся единой семантической сетью для всего текста, использованием числовых значений для узлов и связей, позволяющая реализовать алгоритм формирования структуры естественно-языкового текста для его формализации в информационных системах.
- Предложен алгоритм структурирования естественно-языкового текста для его формализации в информационных системах в соответствии с такими критериями структурированности текста, как наличие иерархического оглавления, наличие заголовков у каждого из разделов оглавления, семантическая связанность внутри каждого из разделов.
- Разработаны методы и алгоритм интеллектуальной обработки текста на основе иерархической и бинарной кластеризации семантической сети для формирования иерархического оглавления естественно-языкового текста, отличающиеся предварительным вычислением агрегирующих характеристик для абзацев и возможностью получать переменное

количество структурных элементов на каждом уровне объединения.

- Разработаны алгоритм интеллектуальной обработки текста на основе методов квазиреферирования для формирования заголовков у каждого из разделов оглавления, позволяющий формировать заголовки из фрагментов сгруппированных частей исходного неструктурированного естественно-языкового текста, делимого по лексемам.
- Разработан алгоритм интеллектуальной обработки текста на основе исчисления предикатов для обеспечения семантической связанности внутри каждого из разделов оглавления, отличающийся использованием правил для предикатов, позволяющих расставить предложения в изначально несвязанных фрагментах естественно-языкового текста в порядке, обеспечивающем наибольшую семантическую связанность получаемого в итоге текста.

Достоверность научных результатов и выводов результатов исследований, полученных автором диссертации, подтверждена строгостью применяемых математических методов и приемлемой степенью согласованности теоретических научных положений с результатами экспериментальных исследований.

Научная и практическая ценность диссертационной работы заключается в том, что результаты, полученные в данной работе, могут быть использованы при обработке неструктурированных текстов, для выделения смысловой нагрузки в учебных и руководящих технических материалах, для определения наиболее актуальных тем при работе RSS-агрегаторов, для педагогических измерительных материалов.

Апробация работы.

Основные положения и результаты диссертации докладывались и обсуждались на 5-й научно-методической конференции «Инновации в науке, образовании и бизнесе» (г. Пенза, 2007 г.), на 14-й научно-методической конференции «Телематика’2007» (г. Санкт-Петербург, 2007 г.), на 15-й научно-методической конференции «Телематика’2008» (г. Санкт-Петербург, 2008 г.) и на научной конференции «Региональная информатика-2008» (г. Санкт-Петербург, 2008 г.)

Публикации.

Основные теоретические и практические результаты диссертации опубликованы в 9 статьях и докладах, из них по теме диссертации 9, среди которых 1 публикация в ведущих рецензируемых изданиях, рекомендованных в действующем перечне ВАК, 3 статьи в других изданиях. Доклады доложены и получили одобрение на 4 международных, всероссийских и межвузовских научно-практических конференциях перечисленных в конце автореферата. Основные положения защищены 1 патентом.

Структура и объем работы.

Диссертация состоит из введения, четырех глав с выводами, заключения. Она изложена на 148 страницах машинописного текста, включает 11 рисунков, 12 таблиц и содержит список литературы из 112 наименований, среди которых 85 отечественных и 27 иностранных авторов.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность исследуемой проблемы, формулируется цель и направление диссертационной работы, приводятся результаты, выносимые на защиту и определяющие научную новизну и практическую ценность результатов исследований.

В первой главе даётся общая характеристика вопросу интеллектуальной обработки текста, в частности рассматривается понятие информационной системы и сущность формализации данных, указывается понятие документа как структурной единицы информационной системы. Далее рассматривается возникновение и развитие области интеллектуальной обработки текста, перечисляются основные направления, существующие в этой области, указываются работы, выполненные в соответствующих направлениях. Отдельно рассматривается классификация моделей естественно-языкового текста. Даётся основная терминология в области интеллектуальной обработки текста, перечисляются основные задачи, решаемые в этой области в настоящее время, кратко указываются их типовые схемы решения. Специально рассматриваются перспективные разработки и направления в области интеллектуальной обработки текста, и указываются нерешённые в настоящее время для этих направлений проблемы, затрудняющие использование соответствующих разработок на практике. Далее делается постановка задачи формирования структуры

изначально неструктурированного естественно-языкового текста, в ходе которой выделяются следующие подзадачи:

- формирование оглавления текста, выявление семантически связанных между собой разделов;
- синтез заголовков для выделенных разделов;
- получение связанного текста внутри выделенных разделов.

Во второй главе проводится анализ существующих моделей естественно-языкового текста на предмет их применимости к решению задачи формирования структуры изначально неструктурированного естественно-языкового текста. Делается вывод о том, что целесообразно использовать модели, получаемые на уровне первичного семантического анализа, далее разработана модель текста, применимая к решению задачи формирования структуры изначально неструктурированного естественно-языкового текста, базирующаяся на модели "смысл-текст" И. А. Мельчука и использующая числовые значения для узлов и дуг единой семантической сети всего текста. Далее выполняется выбор системы первоначальной обработки текста, т. е. преобразования естественно-языкового текста, представленного в виде последовательности символов кодовой таблицы, к используемой модели, и на основе критерия качества первоначальной обработки текста делается вывод о целесообразности использования системы ДИАЛИНГ. Отдельно рассматривается вопрос о выборе конечной формы представления результата, и по соотношению критериев интегрируемости, адапритуемости и операбельности предпочтительным оказывается формат данных, соответствующий одновременно спецификациям HTML и XML.

Затем делается выбор методов решения подзадач, выделенных в первой главе, в ходе которых эти подзадачи в свою очередь делятся на отдельные подзадачи, выделяются отдельные аспекты их решения, для них подбираются соответствующие методы. Первая из подзадач задачи формирования структуры изначально неструктурированного естественно-языкового текста, а именно формирование оглавления текста, разделяется на подзадачи вычисления агрегирующих характеристик для единиц текста и последующей иерархической кластеризации. Для вычисления агрегирующих характеристик за основу берутся статистические данные семантической сети, по сравнению с методами снижения размерности предпочтение отдаётся методам бинарной кластеризации, а именно известному алгоритму fuzzy C-Means, по результатам работы которого подсчитывается значение агрегирующих характеристик для абзацев. Для иерархической кластеризации и формирования оглавления текста выбирается

метод, основанный на предварительной кластеризации известными методами с объединением двух элементов на каждом уровне и последующей перегруппировке с объединением переменного числа элементов на каждом уровне, требующей специального алгоритма. Для синтеза заголовков для выделенных разделов на основе критериев потенциальной эффективности и эффективности существующих реализаций методов реферирования, а также использования в них качества исходного текста, делается выбор в пользу методов квазиреферирования. Для задачи получения связанного текста внутри выделенных разделов делается вывод, что требуются методы, оперирующие небольшими структурными элементами и выполняющие над ними преобразования, что аналогично задаче извлечения знаний из текста и внутреннему представлению знаний, среди методов решения которой известны исчисление предикатов, продукции и фреймы, и в результате сравнения эффективности этих методов в отношении решаемой задачи предпочтительным оказывается использование исчисления предикатов.

В третьей главе в соответствии с выбранными во второй главе методами, выполняется разработка специфических алгоритмов и методов для решения узких подзадач и отдельных их аспектов. Для первой из обозначенных подзадач, которой является вычисление агрегирующих характеристик, разработки специальных алгоритмов не требуется, поскольку основой является известный алгоритм fuzzy C-Means, но требуется разработка специальных методов, а именно:

- выявления понятий, по данным которых выполняется кластеризация;
- приписывания понятиям значений измерений, по которым выполняется кластеризация;
- вычисления агрегирующих характеристик по результатам кластеризации.

За основу исходных данных кластеризации предложено использовать данные о существительных и глаголах, строя матрицу, основывающуюся на числовых значениях дуг, связывающих их в семантической сети. При этом предложено использовать не все лексемы, которым по итогам первоначальной обработки текста были приписаны граммемы глаголов или существительных, а только имеющие наибольшие числовые значения их узлов. На основе нечёткого распределения лексем по кластерам, изначального их распределения по абзацам и числовых значений их узлов предложен метод вычисления агрегирующих характеристик для абзацев.

В задаче иерархической кластеризации и формирования оглавления текста указывается алгоритм преобразования линейного списка абзацев с приписанными им числовыми значениями агрегирующих характеристик в иерархическую структуру, основанный на известных агломеративных методах иерархической кластеризации, а далее разрабатывается алгоритм, выполняющий перегруппировку с объединением переменного числа элементов на каждом уровне при использовании фиксированного количества уровней и заданных для них относительных значений расстояний между центрами кластеров, полученных при исходной иерархической кластеризации.

Для решения задачи синтеза заголовков для выделенных разделов, для которой ранее было предложено использовать метод квазиреферирования, разрабатывается соответствующий алгоритм, выполняющий перебор выделенных в оглавлении разделов и осуществляющий для них два основных действия:

- выбор основных понятий раздела текста для использования их в заголовке;
- поиск фрагмента текста, содержащего оптимальное сочетание выбранных понятий.

Для получения связанного текста внутри выделенных разделов разработаны правила для исчисления предикатов, использующие в качестве входных данных списки понятий и предложений и факты о вхождении понятий в предложения и обеспечивающие на выходе списки, обозначающие расстановку предложений внутри разделов и границы между абзацами. Для упрощения общей программной архитектуры разрабатываемой системы предложено реализовать вывод, применяемый в исчислении предикатов, на алгоритмическом языке программирования с использованием условий и рекурсий.

В четвертой главе составляется концептуальная модель проводимого машинного эксперимента (рис. 1), приводится схема разработанной системы, на основе которой выполняется эксперимент, в виде диаграммы классов (рис. 2), определяется порядок проведения эксперимента применительно к анализу качества решения каждой из выделенных подзадач задачи формирования структуры изначально неструктурированного естественно-языкового текста.

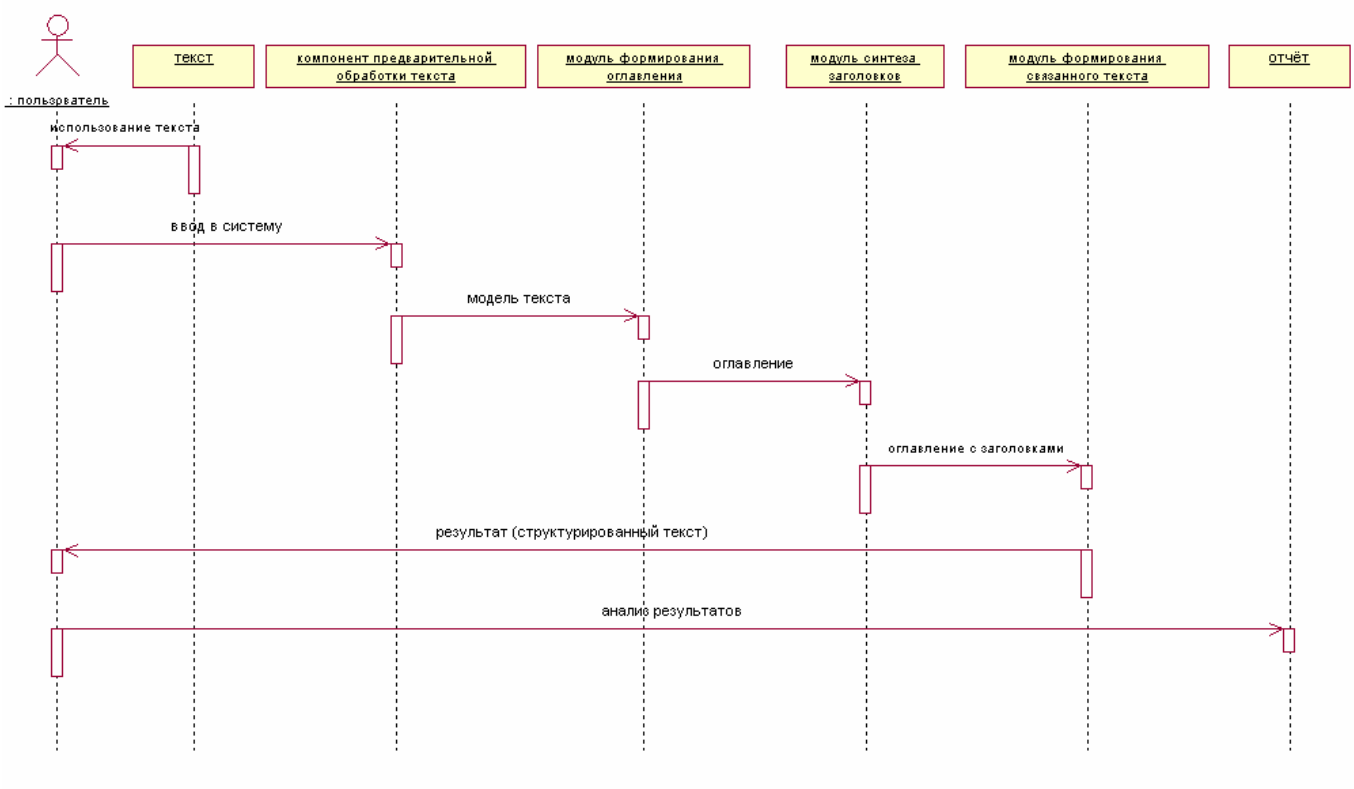

Рисунок 1. Концептуальная модель машинного эксперимента.

Далее даётся более подробное описание экспериментального исследования качества решения каждой из подзадач и приводятся результаты эксперимента. Для подзадачи вычисления агрегирующих характеристик приводится представление результатов в виде графиков и даётся их качественная оценка, для подзадач иерархической кластеризации и формирования оглавления текста, синтеза заголовков для выделенных разделов и получения связанного текста внутри выделенных разделов даётся только качественная оценка полученных результатов.

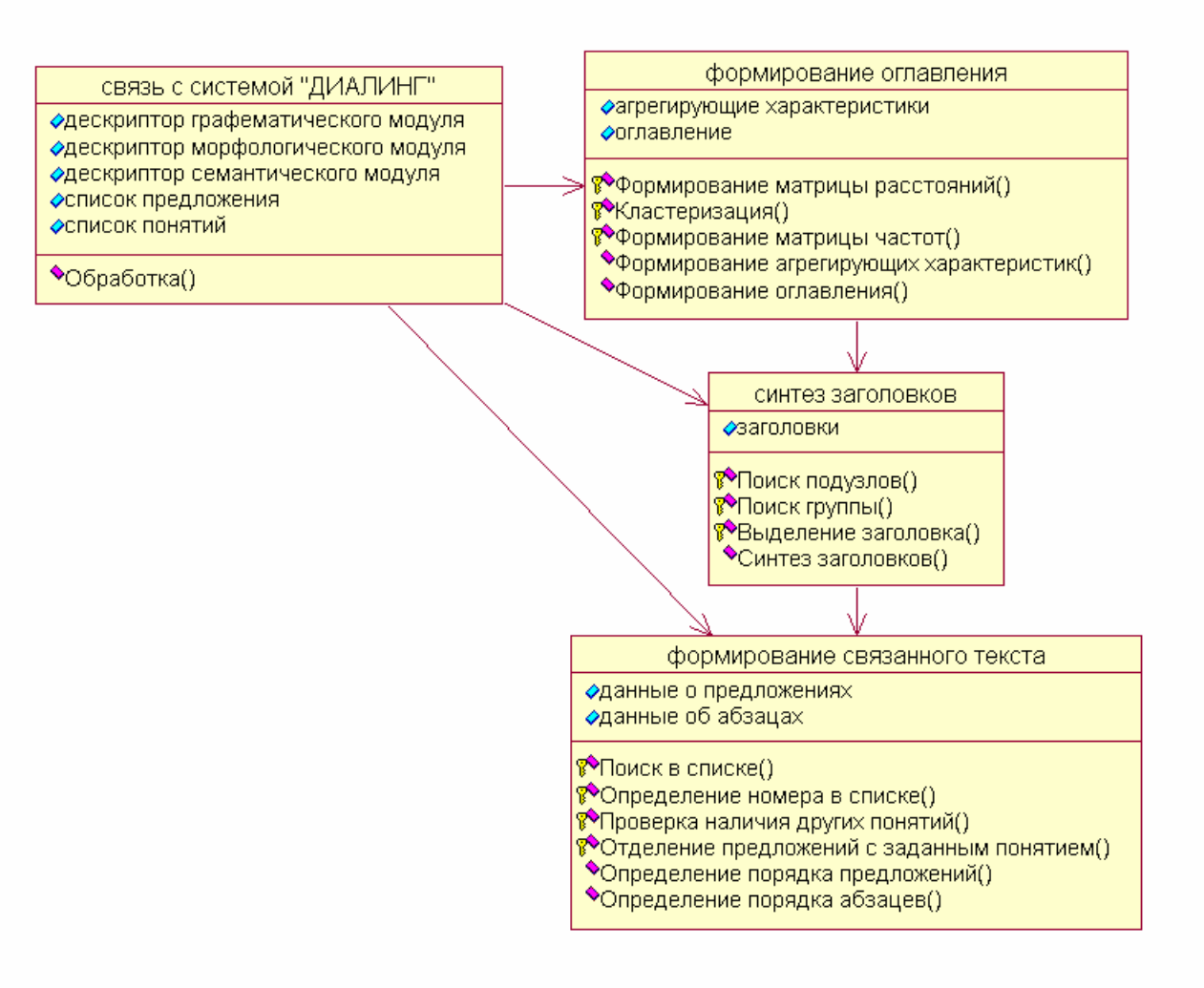

Рисунок 2. Структура разработанной системы.

По результатам эксперимента полученные результаты оказываются удовлетворительными для предварительного формирования структуры изначально неструктурированного естественно-языкового текста при условии последующей их правки вручную.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В работе изложены научно обоснованные новые технические решения по созданию системы структурирования естественно-языкового текста для его формализации в информационных системах, базирующейся на модели его представления первично-семантического уровня, с использованием статистических и лингвистических методов его интеллектуальной обработки.

1. Разработана модель представления естественно-языкового текста, базирующаяся на модели "смысл-текст" в виде семантической сети, отличающаяся единой семантической сетью для всего текста, использованием числовых значений для узлов и связей, что может быть использовано при формировании структуры естественно-языкового текста для его формализации в информационных системах.
2. Разработан алгоритм структурирования естественно-языкового текста для его формализации в информационных системах, отличающийся использованием таких критериев его структурированности, как наличие иерархического оглавления, наличие заголовков у каждого из разделов оглавления, семантическая связанность внутри каждого из разделов.
3. Разработаны методы и алгоритм интеллектуальной обработки текста на основе иерархической и бинарной кластеризации семантической сети для формирования иерархического оглавления естественно-языкового текста, отличающиеся предварительным вычислением агрегирующих характеристик для абзацев и возможностью получать переменное количество структурных элементов на каждом уровне объединения. Бинарная кластеризация выполняется для выделенных понятиям, роль которых в разработанной модели выполняют лексемы, а для их выделение используются весовые значения и граммемы их узлов семантической сети.
4. Разработаны алгоритм интеллектуальной обработки текста на основе методов квазиреферирования для формирования заголовков у каждого из разделов оглавления, позволяющий формировать заголовки из фрагментов сгруппированных частей исходного неструктурированного естественно-языкового текста, делимого по лексемам.

5. Разработан алгоритм интеллектуальной обработки текста на основе исчисления предикатов для обеспечения семантической связанности внутри каждого из разделов оглавления, отличающийся использованием правил для предикатов, позволяющих расставить предложения в изначально несвязанных фрагментах естественно-языкового текста в порядке, обеспечивающем наибольшую семантическую связанность получаемого в итоге текста.
6. Результаты машинного эксперимента показали, что решение задачи структурирования естественно-языкового текста для его формализации в информационных системах отвечает требованиям предварительной обработки при вводе данных в информационную систему при использовании последующей правки вручную. Наилучшие результаты по структурированию естественно-языкового текста для его формализации в информационных системах получены при использовании от 5 до 10 агрегирующих характеристик и пороговой величине весовых значений узлов понятий в тексте от 1 для самых коротких текстов с увеличением на 1 для каждых 4000 символов, а выбранный методов формирования связанного текста существенно не зависит от параметров общего алгоритма.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1. Циликов И. С. Методы и алгоритмы структуризации естественно-языкового текста [Текст] / И.С. Циликов // Системы управления и информационные технологии. № 1.1. (39). 2010. – С. 194-199

Другие статьи и материалы конференций:

2. Пат. № 2008114801 Российская Федерация, МПК7 G01F 17/28. Способ поиска информации в массиве текстов [Текст] / Циликов И. С.; заявитель и патентообладатель Мордов. гос. ун-т. - Заявл. 15.04.2008; опубл. 19.02.2010. 2010

3. Советов Б. Я., Циликов И. С. Вопрос о необходимости стандарта в области Text Mining [Текст] / И.С. Циликов // Труды V Всероссийской научно-методической конференции «Инновации в науке, образовании и бизнесе», Информационно-издательский центр ПГУ, Пенза, 14-15 мая 2007 года

4. Советов Б. Я., Циликов И. С. Комбинированный метод обработки естественных языков [Текст] / И.С. Циликов // Труды XIV Всероссийской научно-методической конференции «Телематика '2007», изд-во ЛИТМО (технический университет), Санкт-Петербург, 19-22 июня 2008 г

5. Циликов И. С. Модель семантики естественного языка [Текст] / И.С. Циликов // Дифференциальная алгебр и динамика систем: Межвуз. сб. науч.-изд-во Мордов. ун-та. 2008, 160 с.. С.131-137

6. Циликов И. С. Подход к выявлению интенсионалов лексем в естественно-языковом тексте [Текст] / И.С. Циликов // Дифференциальная алгебр и динамика систем: Межвуз. сб. науч.-изд-во Мордов. ун-та. 2008, 160 с.. С.137-141

7. Циликов И. С. Подход к решению задачи автоматического построения неявных выводов из естественно-языковых текстов [Текст] / И.С. Циликов // Дифференциальная алгебр и динамика систем: Межвуз. сб. науч.-изд-во Мордов. ун-та. 2008, 160 с.. С.145-149

8. Циликов И. С. Моделирование семантики естественных языков [Текст] / И.С. Циликов // Труды XV Всероссийской научно-методической конференции «Телематика-2008», Санкт-Петербург, 23-26 июня 2008 г

9. Циликов И. С. Метод применения правил формальных грамматик для глубинного семантического анализа [Текст] / И.С. Циликов // Материалы XI Санкт-Петербургской международной конференции «Региональная информатика-2008 «РИ-2008», Санкт-Петербург, 21-24 октября 2008 г